2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

LIDIA: Localizing In the Dark with Illumination-Awareness toward Perception-Aware Planning

Iason Georgios Velentzas[1] and Kento Tomita[2]

***Abstract*— Accurate Localization is a fundamental challenge in robotic autonomy, with applications ranging from autonomous driving to space proximity operations. Visual Localization is a viable choice in GPS-denied environments, such as subterranean, indoor, urban, or space environments; however, its performance degrades under often encountered conditions, such as low light or varying illumination. This paper introduces LIDIA — an illumination-aware model of localization quality for Perception-Aware Planning. LIDIA involves the efficient integration of light source direction into the planning framework, enabling the prediction of visually informative regions in the map under varying lighting. Unlike prior geometric approaches, LIDIA jointly exploits geometric and photometric information without requiring computationally expensive real-time rendering, thereby preserving online applicability. Our results demonstrate that LIDIA consistently outperforms existing geometric methods such as FIF in predicting the information gain of candidate camera poses and in planning trajectories that achieve higher localization accuracy. To the best of our knowledge, this is the first approach to unify geometric and photometric reasoning in an efficient, active localization system, paving the way for robust autonomy in illumination-constrained environments.**

***Index Terms*— Visual localization, Perception-aware Planning, Active SLAM, Illumination-awareness, Robot autonomy**

I. Introduction

A fundamental requirement for achieving higher levels of autonomy in robotic systems is the ability to accurately localize within a preexisting Map of the environment. Visual Localization (VLOC) has emerged as a solution to localizing in environments where GPS measurements are not reliable, such as urban, indoor, subterranean, and space or planetary exploration environments. Since VLOC is predicated on the ability of a robot to process visual input, researchers realized early on the coupling between perception and motion planning — often termed *Perception-aware Planning*, or *active perception*. The concrete objective depends on the goal of the application, ranging from minimizing mapping time to maximizing localization accuracy within a Map. This work focuses on active VLOC by trying to efficiently and accurately model the localization quality of arbitrary poses. Then, the planning module can ensure that sufficient localization information is present throughout the designed trajectory.

[1]Iason Georgios Velentzas is with School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA iason.velentzas@gatech.edu. This work was developed during the author's internship at MERL.

[2]Kento Tomita is with Mitsubishi Electric Research Laboratories, Cambridge, MA, 02139, USA tomita@merl.com

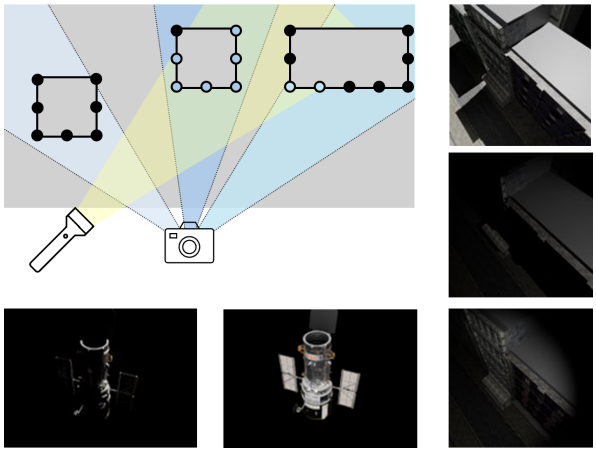

Fig. 1. The top left schematic depicts three distinct camera orientations and a specified lighting direction, where some landmark points are visible, but not illuminated. The right column and bottom row depict different Sun directions of the same viewpoint of an urban outdoor environment and a spacecraft in orbit, respectively.

Notably, the environments where VLOC is needed suffer from harsh illumination conditions, either partially — indoor and outdoor facilities during the night with insufficient illumination —, or completely — subterranean and space environments with a localized varying light source. The illumination conditions are one of the primary sources of performance degradation in VLOC systems, heavily affecting the appearance of the images, and hence the localization accuracy. Examples of this strong effect are illustrated in Fig. 1. The right column depicts an urban outdoor environment during the night under different orientations of the flashlight, while the bottom row shows a spacecraft in orbit under different sun directions. Current approaches estimate the localization quality of the pose using only geometric information and essentially assuming that geometrically visible landmarks will always contribute to localization quality. However, this is not the case in real-world applications where the illumination is varying, severely affecting the appearance of the images.

To alleviate this assumption, we propose **LIDIA** — **L**ocalizing **I**n the **D**ark with **I**llumination **A**wareness — an illumination-aware framework that uses an estimate of the light source direction to combine geometric and photometric reasoning about the contribution of the Map landmarks to the localization module. A 2D example of the configuration is shown in Fig. 1. We demonstrate how to efficiently integrate photometric information in Perception-aware Planning and enhance the fidelity of the localizability model, without the need for expensive real-time rendering. We evaluate LIDIA

by running a complete perception pipeline on synthetic images under varying illumination conditions that assess the applicability and robustness of our algorithm in real-life missions. We demonstrate how the incorporation of the light source direction information leads to significantly more accurate prediction of the localization quality of poses, as well as improved planning decisions that lead to higher localization accuracy.

II. Related Work

VLOC involves the estimation of the pose of the camera sensor from a given image and knowledge about the environment. Image retrieval methods store image databases instead of a 3D Map and retrieve images from the database that are similar to the query image. NetVLAD [1] or DenseVLAD [2] are widely used to select candidate images from which an initial pose is estimated by interpolating poses from the database. Abolute Pose Regression (APR) techniques regress the camera pose directly from a single image. Many works utilize Deep Neural Netowrks for one-shot pose estimation [3]–[5], while Sattler et al. discuss some limitations of these techniques [6]. Scene Coordinate Regression (SCR) methods such as DSAC [7] try to estimate the 3D positions of each pixel and then regress the pose via optimization blending both learning and traditional techniques. SCR however, require dense pixel-wise supervision with ground-truth 3D scene coordinates, which is hard to obtain in many cases.

With the emergence of more accurate novel view synthesis techniques, such as NeRF [8] or 3DGS [9] there have been applications of this field for the VLOC task [10], [11]. A limiting factor is that these techniques are inherently appearance-based and are not preserving geometric consistency, while at the same time the rendering requirements are not compatible with the needs for fast evaluation of the planning module.

Finally, structure-based localization methods remain the most accurate and robust across diverse scenarios. Early works relied on associating 2D features with 3D landmarks in the Map [12], [13] followed by a pose optimization step. More recent approaches incorporate ideas from image retrieval [14] or adopt hierarchical localization strategies [15].

The methods mentioned before are inherently passive — they operate on the available observations. However, if the robot is able to predict the quality of an observation, it can influence its pose to achieve its goal. This coupling of planning and perception created the field of Perception-aware planning, where the robot seeks to actively design its trajectory to affect localization. Early Active Localization approaches used discrete Bayesian models to greedily select actions that reduce pose uncertainty [16], [17]. A significant portion of the literature is focused on improving geometric-based information on pose uncertainty with heuristics. Zhang et.al. [18] evaluate a library of candidate trajectories to safely reach a target location combining heuristics for perception quality, collision probability, and distance to the goal location. Coupling the visibility of landmarks with their semantic information, the authors of [19] propose to steer the camera towards the semantically most informative regions.

Dealing with the considerations of employing algorithms in real-life scenarios, the Fisher Information Field (FIF) [20] was proposed to evaluate candidate poses with constant query time, by pre-computing geometrically unoccluded landmarks on a 3D grid world. The authors proposed to use Gaussian processes for field-of-view visibility estimation, which results in a smooth localization quality model that can be used with continuous optimization techniques. Inspired by FIF, the recent work of [21] is a self-supervised data-driven alternative to select informative viewpoints, allowing greater flexibility in viewpoint selection and maintaining real-time operation.

Despite the fact that the research community has explored the coupling between perception and planning, most of these works assume stable illumination conditions. In practice, however, illumination can vary drastically — from shadows and low-light settings to strong specular reflection effects — and degrade the performance of the localization pipeline. Only a relatively small part of literature deals with varying illumination, and it is mainly devoted to image enhancement and relighting. Applications in Localization and Mapping typically preprocess the input image with Gamma correction [22], Retinex Theory [23], or with GAN-assisted image enhancement [24]. More recently, the authors of [25], [26] propose learning relighting parameters along with the 3D Gaussian representation, in order to generate visually consistent images with different illumination conditions.

In summary, Perception-aware Planning frameworks typically overlook the photometric effects of dynamic illumination. LIDIA addresses the gap in the literature by explicitly modeling the lighting conditions and unifying geometric visibility and photometric information to predict the localization quality of a pose.

III. Preliminaries

In this section, we detail the point-based VLOC pipeline and the per-landmark Fisher Information Matrix derivation. We define the processes of feature detection, matching, and pose estimation using a known 3D landmark map.

A. Feature Detection and Description

Consider a query image $I : \Omega \to \mathbb{R}$ and a reference image $I' : \Omega \to \mathbb{R}$ with pixel domain $\Omega \subset \mathbb{R}^2$. The feature detector localizes *keypoints* $u_k, u'_k \in \Omega$ in the two images and then the descriptor produces a d-dimensional descriptor vector $\mathbf{d}_k \in \mathbb{R}^d$ for each keypoint that describes the image around the keypoint. In total, the detection and description process $\mathcal{D}$ applied to each image produces the following.

$$\mathcal{D}(I) = \{(u_i, \mathbf{d}_i)\}_{i=1}^{N}, \qquad \mathcal{D}(I') = \{(u'_j, \mathbf{d}'_j)\}_{j=1}^{N'} \quad (1)$$

B. Feature Matching

The feature description process should ideally describe corresponding keypoints with similar descriptors. The feature matching process seeks to associate the sets of descriptors

between the two images. This can be done by deep neural architectures or traditionally through a mutual nearest-neighbor association.

The outcome of feature matching is the keypoint correspondences $\mathcal{M} = \{(\mathbf{u}_i, \mathbf{u}'_j) \mid i \in \mathcal{J} \subseteq \{1, \dots, N\}, j \in \mathcal{J}' \subseteq \{1, \dots, N'\}\}$, which can be subsampled to an inlier set using the epipolar constraint and Fundamental Matrix estimation [27]:

$$\hat{\mathcal{M}} = \{(\mathbf{u}_i, \mathbf{u}'_j) \mid i \in \hat{\mathcal{J}} \subseteq \mathcal{J}, j \in \hat{\mathcal{J}}' \subseteq \mathcal{J}'\} \tag{2}$$

C. Pose Estimation

The typical process of pose estimation on a landmark-based Map involves running PnP [28] with RANSAC on the 2D-3D correspondences. For this work, we assume that there exists a Map $\mathcal{X}$ with landmarks $X_l^{\mathrm{w}} \in \mathbb{R}^3, \quad l = 1, \dots, L$ and that the *reference* image has already known associations with the landmark Map:

$$\mathcal{A}_{\mathrm{ref}} = \{(\mathbf{u}'_j, X) \;\mid\; \mathbf{u}'_j \in \mathcal{D}(I') \;\wedge\; X \in \mathcal{X}\} \tag{3}$$

We then utilize the inlier matches between the *query* and the *reference* images to associate the keypoints of the *query* image and the Map:

$$\mathcal{A}_{\mathrm{q}} = \{(\mathbf{u}_i, X) \;\mid\; (\mathbf{u}_i, \mathbf{u}'_j) \in \hat{\mathcal{M}} \wedge (\mathbf{u}'_j, X) \in \mathcal{A}_{ref}\} \tag{4}$$

Let the relative pose $\mathbf{T}_{\mathrm{cw}} \in \mathrm{SE}(3)$ of the camera frame $\mathcal{C}$ with respect to the Map frame $\mathcal{W}$ be

$$\mathbf{T}_{\mathrm{cw}} = \begin{bmatrix} \mathbf{R}_{\mathrm{cw}} & \mathbf{t}_{\mathrm{wc}}^{\mathrm{c}} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \; \mathbf{R}_{\mathrm{cw}} \in \mathrm{SO}(3), \; \mathbf{t}_{\mathrm{wc}}^{\mathrm{c}} \in \mathbb{R}^3. \tag{5}$$

and a point $X \in \mathbb{R}^3$ of the Map have homogeneous coordinates $\bar{X} = (X^\top, 1)^\top$. Using the standard pinhole camera projection matrix, [29] with focal length $f_x, f_y \in \mathbb{R}$ and principal points $c_x, c_y \in \mathbb{R}$ in the resepctive image axes, we define the forward-projection function of a landmark X in the camera plane to be

$$\Pi(\mathbf{T}_{\mathrm{cw}}, X) = \pi(\mathcal{K} \begin{bmatrix} \mathbf{R}_{\mathrm{cw}} \mid \mathbf{t}_{\mathrm{wc}}^{\mathrm{c}} \end{bmatrix} \bar{X}), \;\; \pi\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \begin{bmatrix} x/z \\ y/z \end{bmatrix}, \tag{6}$$

where K is the camera calibration matrix. So, given a 2D-3D correspondence $(\mathbf{u}_i, X) \in A_q$ and a relative pose $\mathbf{T}_{\mathrm{cw}}$, the reprojection error is defined as

$$\varepsilon(\mathbf{u}_i, X, \mathbf{T}_{\mathrm{cw}}) = \|\mathbf{u}_i - \Pi(\mathbf{T}_{\mathrm{cw}}; X)\|_2. \tag{7}$$

Given a pixel threshold ϵ_{PnP}, the inlier set for $\mathbf{T}_{\mathrm{cw}}$ is

$$\mathcal{P}(\mathbf{T}_{\mathrm{cw}}) = \big\{ (\mathbf{u}_i, X) \in \; \mathcal{A}_q \mid \varepsilon(\mathbf{u}_i, X, \mathbf{T}_{\mathrm{cw}}) < \epsilon_{\mathrm{PnP}} \big\}. \tag{8}$$

RANSAC samples iteratively subsets of correspondences to solve for a pose estimate and selects the one with the most inliers across the full correspondence set:

$$\mathbf{T}_{\mathrm{cw}}^{\mathrm{ransac}} = \arg\max_{\mathbf{T}_{\mathrm{cw}}} \; |\mathcal{P}(\mathbf{T}_{\mathrm{cw}})|. \tag{9}$$

Let $\mathcal{P}^* = \mathcal{P}(\mathbf{T}_{\mathrm{cw}}^{\mathrm{ransac}})$ be the chosen inlier indices. The final pose is obtained by refitting only on these inliers, which solely contribute to the final localization accuracy.

D. Per-Landmark Fisher Information Matrix

To calculate the Fisher Information Matrix of a landmark, we use a bearing-based measurement model, similarly to previous work [20]. The bearing vector associated with a landmark X^{w} measured from a relative camera pose $\mathbf{T}_{\mathrm{cw}}$ is:

$$f = \frac{X^{\mathrm{c}}}{\|X^{\mathrm{c}}\|}, \qquad X^{\mathrm{c}} = R_{\mathrm{cw}}\, X + t_{\mathrm{wc}}^{\mathrm{c}} \in \mathbb{R}^3 \tag{10}$$

A world-frame pose perturbation $\delta\xi^w = (\delta\rho^w, \delta\omega^w) \in \mathbb{R}^6$ induces the following first-order change:

$$\frac{\partial p_c}{\partial \xi^{\mathrm{w}}} \;=\; R_{cw} \begin{bmatrix} -I_3 & [X^{\mathrm{w}}]_\times \end{bmatrix} \;\in \mathbb{R}^{3\times 6}, \tag{11}$$

where $[X]_\times$ is the skew matrix of X. The Jacobian of the bearing w.r.t. X^{c} is

$$\frac{\partial f}{\partial X^{\mathrm{c}}} \;=\; \frac{1}{\|X^{\mathrm{c}}\|} \left(I_3 - f f^\top\right) \;\in\; \mathbb{R}^{3\times 3}. \tag{12}$$

Combining (11)–(12), the Jacobian measurement per landmark w.r.t. the global pose is

$$J(\mathbf{T}_{\mathrm{cw}}; X^{\mathrm{w}}) \;=\; \frac{\partial f}{\partial X^{\mathrm{c}}}\, R_{cw} \begin{bmatrix} -I_3 & [X]_\times \end{bmatrix} \;\in\; \mathbb{R}^{3\times 6}. \tag{13}$$

We assume an isotropic measurement covariance $\sigma^2 I_3$ in bearing space with $\sigma^2 > 0$, which leads to the formulation of the per-landmark Fisher information Matrix (FIM):

$$\mathrm{FIM}(\mathbf{T}_{\mathrm{cw}}, X^{\mathrm{w}}) = \sigma^{-2} J^\top J, \qquad J = J(\mathbf{T}_{\mathrm{cw}}, X^{\mathrm{w}}). \tag{14}$$

IV. Problem Statement

The objective of this work is to predict the localization quality of a relative pose with respect to a landmark Map, *without* relying on the image that would be captured from this pose. Suppose that we are given the setup of a camera calibration matrix $\mathcal{K}$, and a pre-built sparse collection of landmarks, the Map of the environment, expressed in the world frame $\mathcal{X} = \{X_l^w \in \mathbb{R}^3, l = 1, \cdots, L\}$. Moreover, assume that we are given as inputs the camera extrinsic matrix $\mathbf{T}_{\mathrm{cw}} \in \mathrm{SE}(3)$, and the type of light sources t_ℓ operating in the environment along with their pose w.r.t the world frame $\mathcal{L} = \{\ell_i = (t_{\ell_i}, \mathbf{T}_{\ell_i \mathrm{w}}), i = 1, \dots, n\}$. The light source input is typically not part of perception-aware planning algorithms, since the approaches are either purely geometric, or they assume similar illumination conditions with the mapping process. However, in lighting-constrained environments, there is a critical dependence between the light sources, and hence the image appearance, and the localization quality of a pose.

Treating localization as a parameter estimation problem, where we seek to estimate the pose from which an image is captured, a good localization quality metric is the Cramer-Rao lower bound, which quantifies the lowest covariance attainable by an impartial estimator [30] and is calculated through the inverse of FIM. As a result, we can use the IG of each landmark to quantify localization quality as the aggregate information gain of the landmarks belonging to the PnP inlier set of the perception pipeline, since only these landmarks contribute to the final pose estimation. Notice that

during prediction, the final PnP inlier set is unknown and needs to be estimated.

Under this formulation, we explicitly calculate the ideal quantity to be predicted. The image captured from $\mathbf{T}_{\text{cw}}$ can be processed through the perception pipeline to obtain a pose estimate $\mathbf{T}^{\star}_{\text{cw}}$ with respect to the map and the corresponding inlier landmark set $\mathcal{P}^{\star}$. As a result, the actual localization quality of the pose is as follows:

$$\mathcal{Q}(\mathcal{P}^{\star}) = \sum_{X^{\text{w}} \in \mathcal{P}^{\star}} \text{IG}(\mathbf{T}_{\text{cw}}, X^{\text{w}}) \tag{15}$$

Problem. Given the setup configuration $(\mathcal{X}, \sigma^2, \mathcal{K})$ and the inputs $(\mathbf{T}_{\text{cw}}, \mathcal{L})$, we wish to define an estimator of the localization quality of the configuration, *without* the associated image, which essentially means estimating the latent inlier set of the actual perception pipeline $\mathcal{P}_{\text{LIDIA}}(\mathbf{T}_{\text{cw}}, \mathcal{L})$, such that:

$$\mathcal{Q}(\mathcal{P}_{\text{LIDIA}}(\mathbf{T}_{\text{cw}}, \mathcal{L})) \approx \mathcal{Q}(\mathcal{P}^{\star}) \tag{16}$$

V. Methodology

Geometric visibility has been used extensively, but almost exclusively in the literature to predict the landmarks that are contributing to localization. A landmark is considered to be geometrically visible if it is unoccluded and inside the Field of View (FoV) of the camera. The position of the camera defines the occlusion, while the camera orientation defines whether the landmark resides in the FoV.

Different techniques have been used in the literature to calculate geometric visibility. The authors of FIF [20] exploit depth measurements to determine the occlusion visibility of landmarks and Gaussian Process Regression to learn the FoV visibility. In [21], the authors opt for a z-buffer algorithm when dense depth data is available or for the Hidden Point Removal Operator, when it is not. Finally, from a more theoretical perspective, the authors of [31] assume a visibility-weighted landmark distribution on the image.

In this work, we propose to approximate the latent inlier set of landmarks, $\mathcal{P}^{\star}$, by the set of landmarks that are both geometrically visible and directly illuminated from the light sources.

A. Occlusion Visibility.

We employ ray-casting from the camera position towards the landmarks to determine occlusion visibility. We use the dense scene mesh $\mathcal{S}$ reconstructed from Structure-from-Motion [32]. We define the direction from the camera position $\mathbf{t}^{\text{w}}_{\text{cw}}$ to the landmark X:

$$\mathbf{v}(\mathbf{t}^{\text{w}}_{\text{cw}}, X) = X - \mathbf{t}^{\text{w}}_{\text{cw}} \tag{17}$$

and then the line segment ray-cast between these two points can be described with parameter $\tau \in \left[0,\ \|\mathbf{v}(\mathbf{t}^{\text{w}}_{\text{cw}}, X)\|\right]$ as:

$$r(\tau; \mathbf{t}^{\text{w}}_{\text{cw}}, X) \ = \ \mathbf{t}^{\text{w}}_{\text{cw}} + \tau\, \hat{\mathbf{v}}(\mathbf{t}^{\text{w}}_{\text{cw}}, X), \tag{18}$$

where $\hat{\mathbf{v}}$ denotes the unit direction of $\mathbf{v}$. This ray-cast creates the set of unoccluded landmarks:

$$\mathcal{P}_{\neg\,\text{occ}}(\mathbf{t}^{\text{w}}_{\text{cw}}) = \{X \in \mathcal{X} \mid r(\tau; \mathbf{t}^{\text{w}}_{\text{cw}}, X) \cap \mathcal{S} = \varnothing\}. \tag{19}$$

B. FoV Visibility.

We check whether the forward projection of the landmark lies inside the image domain, so with an image domain Ω, this yields:

$$\mathcal{P}_{\text{FoV}}(\mathbf{T}_{\text{cw}}) = \Big\{X \in \mathcal{X} \ \Big|\ \Pi(\mathbf{T}_{\text{cw}}; X) \in \Omega\Big\}. \tag{20}$$

Hence, the set of geometrically visible landmarks is:

$$\mathcal{P}_{\text{geo}}(\mathbf{T}_{\text{cw}}) = \mathcal{P}_{\neg\ \text{occ}}(\mathbf{t}^{\text{w}}_{\text{cw}}) \cap \mathcal{P}_{\text{FoV}}(\mathbf{T}_{\text{cw}}). \tag{21}$$

C. Illumination Awareness.

The set of directly illuminated landmarks is a subset of illuminated landmarks, in general. When there is strong indirect illumination from complex geometry and reflections in the scene, the set of directly illuminated landmarks is a conservative approximation of the illuminated landmarks. However, this approximation should be sufficient for planning the camera poses, since in general, the directly illuminated landmarks are expected to survive at higher rates through the perception pipeline when compared to indirectly illuminated ones.

1) Sun Light Source: We model each light source as $\ell = (\theta, \mathbf{T}_{\text{wl}})$, where θ is the half-angle of the emission cone of the light source, while $\mathbf{T}_{\text{wl}}$ is the relative pose of the light source w.r.t the world frame. From the relative pose, we can retrieve the position of the light source in the world frame $\mathbf{t}^{\text{w}}_{\text{lw}}$ and the direction the light is being shed is the z-axis of the light frame: $\ell^{\text{w}}_{\text{d}} = \mathbf{R}_{\text{wl}}\mathbf{e}_z \in SO(2)$, where $\mathbf{e}_z = [0, 0, 1]^{\top}$. The emission angle defines the type of light source, since flashlights are modeled with $\theta \in (0, \pi]\, rad$, while sun is arbitrarily modeled as $\theta = 0$.

2) Flashlight Light Source: The flashlight source is emulated using a conic spotlight, where the emission cone can be viewed as the FoV of the light source. The directly illuminated landmarks must be unoccluded from the light source and inside the emission cone. Illumination occlusions for a flashlight, without considering the emission cone, are calculated similarly to the geometric ones using ray-casting:

$$\mathcal{P}^{\text{illum}}_{\neg\,\text{occ}}(\mathbf{t}^{\text{w}}_{\text{lw}}) = \{X \in \mathcal{X} \mid r(\tau; \mathbf{t}^{\text{w}}_{\text{lw}}, X) \cap \mathcal{S} = \varnothing\}. \tag{22}$$

The set of landmarks that are inside the emission cone without considering occlusions is:

$$\mathcal{P}^{\text{illum}}_{\text{FoV}}(\mathbf{T}_{\text{lw}}, \theta) = \Big\{X \in \mathcal{X} \ \Big|\ \cos^{-1}\left(\hat{\mathbf{v}}(\mathbf{t}^{\text{w}}_{\text{lw}}, X), \ell^{\text{w}}_{\text{d}}\right) \leq \theta\Big\}. \tag{23}$$

As a result, the final set of directly illuminated points is:

$$\mathcal{P}_{\text{illum}}(\ell) = \mathcal{P}^{\text{illum}}_{\neg\text{occ}}(\mathbf{t}^{\text{w}}_{\text{lw}}) \cap \mathcal{P}^{\text{illum}}_{\text{FoV}}(\mathbf{T}_{\text{lw}}, \theta). \tag{24}$$

We model the Sun as a directional light, a point source at infinity, producing spatially uniform, parallel rays towards the $\ell^{\text{w}}_{\text{d}}$ direction. We employ inverse ray-casting from the landmarks towards $-\ell^{\text{w}}_{\text{d}}$ to identify directly illuminated landmarks. Using a sufficiently large scene diameter $D_{\mathcal{X}}$, we define the inverse ray-cast line segments from each landmark as:

$$r(\tau; \ell^{\text{w}}_{\text{d}}, X) \ = \ X - \tau\, \ell^{\text{w}}_{\text{d}}, \qquad \tau \in \left[0, D_{\mathcal{X}}\right]. \tag{25}$$

The directly illuminated points are the ones that are illuminated unobstructed from the specified ray-casts:

$$\mathcal{P}_{\text{illum}}(\ell) = \{X \in \mathcal{X} \mid r(\tau; \ell_{\text{d}}^{\text{w}}, X) \cap \mathcal{S} = \varnothing\} \tag{26}$$

D. Localization Quality

We create an approximate model of the final latent inlier set of landmarks that considers the landmarks that are both geometrically visible from the camera and directly illuminated by at least one of the light sources:

$$\mathcal{P}_{\text{LIDIA}}(\mathbf{T}_{cw}, \mathcal{L}) = \mathcal{P}_{\text{geo}}(\mathbf{T}_{cw}) \bigcap \left(\bigcup_{\ell \in \mathcal{L}} \mathcal{P}_{\text{illum}}(\ell) \right) \tag{27}$$

Hence, the localization quality of the pose is estimated as:

$$\mathcal{Q}\left(\mathcal{P}_{\text{LIDIA}}(\mathbf{T}_{cw}, \mathcal{L})\right) \tag{28}$$

VI. Experiments

The purely geometric counterpart to our approach is the FIF method [20]. Since our formulation does not incorporate real-time performance enhancements, we conducted the comparison without employing the Gaussian Process regression and 3D voxel grid approximations introduced in FIF. We evaluated against the exact formulation, which reduces to the summation of per-landmark Fisher Information Matrices (FIM), as defined in (14). The trace of FIM is employed to convert it to a scalar quantity, while the perception pipeline remains consistent across the two approaches.

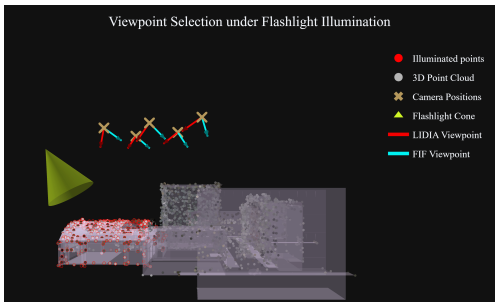

Fig. 2. FIF and LIDIA optimal camera orientations for fixed camera positions and sun direction in an urban environment.

We evaluated the performance of LIDIA through a series of qualitative and quantitative experiments with synthetic images generated in Blender [33]. The 3D landmark Map is generated using COLMAP [32] on a set of images with favorable illumination, which is maintained as the database. For the mapping process, we use NetVLAD [1] for image association, Superpoint [34] for feature detection, and Superglue [35] for feature matching. To measure the accuracy of the localization quality estimation by LIDIA, along with the inlier set $\mathcal{P}_{\text{inlier}} = \mathcal{P}^{\star}$, we define the sets of landmarks corresponding to the detected and matched features.

$$\mathcal{P}_{\text{det}} = \{X \in \mathcal{X} \mid (u, \mathbf{d}) \in \mathcal{D} \wedge \varepsilon(u, X, \mathbf{T}) < \epsilon_{\text{PnP}}\} \tag{29}$$

$$\mathcal{P}_{\text{mat}} = \{X \in \mathcal{X} \mid (u, u') \in \hat{\mathcal{M}} \wedge \varepsilon(u, X, \mathbf{T}) < \epsilon_{\text{PnP}}\}, \tag{30}$$

where $\mathcal{X}$ is the reconstructed map, $\mathbf{T}$ is the relative camera pose, and $\mathcal{D}$, $\hat{\mathcal{M}}$ denote detection and matching on the query and reference images, respectively.

Optimal view directions are selected by maximizing the localization quality $\mathcal{Q}$ estimated by each method, LIDIA and FIF. For maximization, we employ sampling-based grid search on the unit sphere centered at the designated camera positions with a Fibonacci lattice.

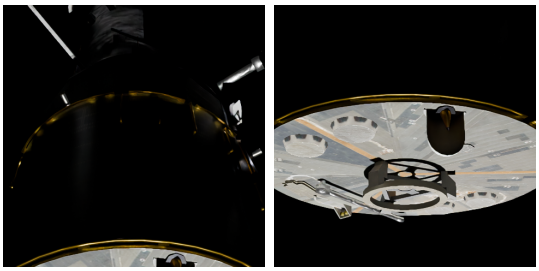

Fig. 3. Different viewpoint selections from the same camera position of Closer range for FIF (left) and LIDIA (right).

A. Qualitative Result in Urban Scene

As an initial analysis, we compared the estimated most informative view directions based on LIDIA and FIF in an urban scene with a fixed conic flashlight. Fig. 2 confirms that LIDIA points toward illuminated regions whereas FIF looks towards areas closer to the camera position even if they are not illuminated. Similarly, in a spacecraft approach scenario in Fig. 3, we see that LIDIA selects the illuminated lid over the dark body of the spacecraft, which will probably contain more texture information.

B. Quantitative Results in Space Scene

A representative scenario where harsh illumination affects localization quality is the spacecraft scene. We reconstructed a landmark map of the Hubble Space Telescope (HST) [36] and analyzed the localization quality for two distinct ranges of camera positions and multiple sun directions, as depicted in Fig. 4. The **Database range** corresponds to images captured from distances comparable to the database images, where a substantial portion of the spacecraft is visible. The **Closer range** corresponds to images captured only a few meters away from the spacecraft surface, where only small portions of the spacecraft are visible.

TABLE I
Localization performance in optimized view directions; Database range (left) and Closer range (right).

Loc Acc ↑	FIF	LIDIA
5cm, 0.4°	5.7%	**14.4%**
0.25m, 2°	23.9%	**41.5%**
0.5m, 3°	32.9%	**49.4%**
1m, 5°	39.8%	**52.8%**

Loc Acc ↑	FIF	LIDIA
5cm, 0.4°	4.5%	**20.5%**
0.25m, 2°	11.4%	**43.2%**
0.5m, 3°	18.2%	**70.5%**
1m, 5°	40.9%	**90.5%**

For Database range images, the main differences with respect to the database images are viewpoint and illumination, enabling the use of the database images for localization with image retrieval. We use the NetVLAD descriptor to select the 10 best candidate images from the database, from which the 2D-3D correspondences are concatenated.

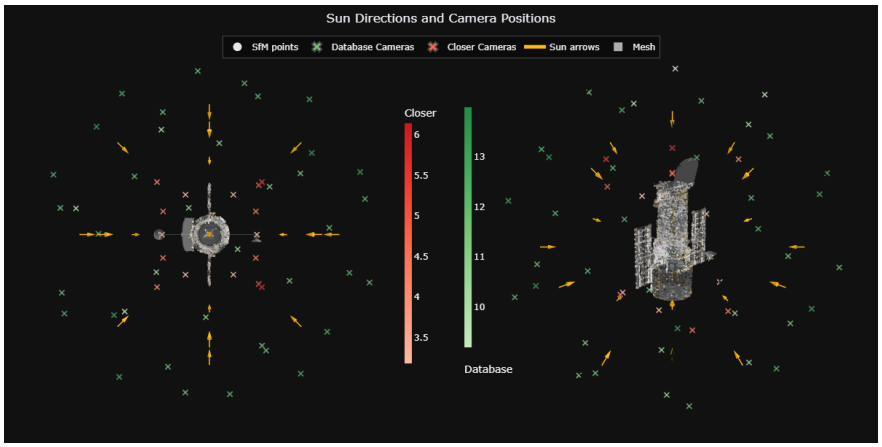

Fig. 4. Two different viewpoints of the camera positions and sun directions for Database (Green) and Closer (Red) range query images.

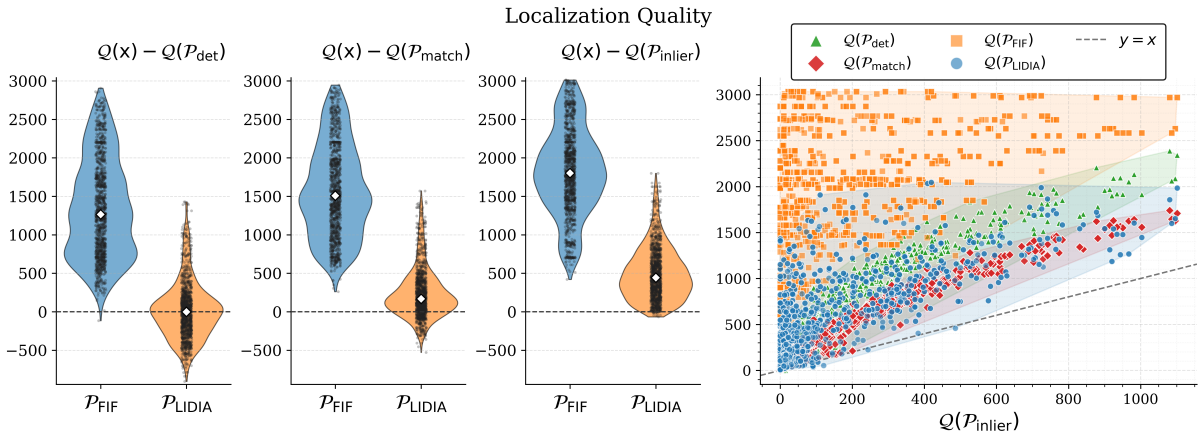

Fig. 5. **Database range:** The left-most panels show the Information Gain (IG) difference between the predicted estimates and the detected, matched, and actual IG, respectively. The right panel shows the IG as a function of the actual IG.

In contrast, Closer range images additionally capture the spacecraft at a different scale, which significantly degrades the performance of image retrieval techniques. To address this, we generated reference images from randomly selected nearby viewpoints, ensuring sufficient visual overlap with each query image. This setup maintains consistent illumination and scale between query and reference images, resulting in a SLAM-like evaluation appropriate for this scenario. In this case, we use ground-truth knowledge about the rendered reference images to establish 2D-3D correspondences.

We report the distribution of the difference between the LIDIA or FIF localization quality estimates and each of the quantities $\mathcal{Q}(\mathcal{P}_{\text{det}})$, $\mathcal{Q}(\mathcal{P}_{\text{match}})$, and $\mathcal{Q}(\mathcal{P}_{\text{inlier}})$. We also investigated the localization error metrics of the FIF- and LIDIA-selected poses, as adapted from the Long-Term Visual Localization benchmark [37].

1) Database range: Fig. 5 reports the differences per image in the predicted IG relative to the detected, matched, and actual IG. It is evident from the plots that LIDIA maintains a tighter spread around all reference values. It is also confirmed that LIDIA underestimates the detected IG, since it assumes that landmarks not directly illuminated are not visible, thus providing a conservative estimate. However, we see that as IG is processed through the pipeline towards the actual IG, the predictions of LIDIA remain in bulk close to the reference values. On the other hand, it is clear that as we move towards actual IG, the FIF predictions deviate from the reference values. The right panel of Fig. 5 presents a scatterplot of the localization quality per image. We notice that LIDIA follows the trend of the reference values, while FIF remains uncorrelated with the actual IG.

The left part of Table I verifies that a better estimate of the localization quality results in selecting better viewpoints in terms of localization accuracy. We observe a significant and consistent increase in the percentage of query images that are localized throughout every threshold.

2) Closer range: Fig. 6 exhibits patterns similar to the case of the Database range. However, LIDIA appears more conservative with respect to detected IG, as the Closer range experiment operates at a scale different from that of the Map reconstruction. Even though the existing finer details in these images lead to more indirectly illuminated landmarks being detected, these landmarks do not seem to survive throughout the Localization pipeline. LIDIA is predicated

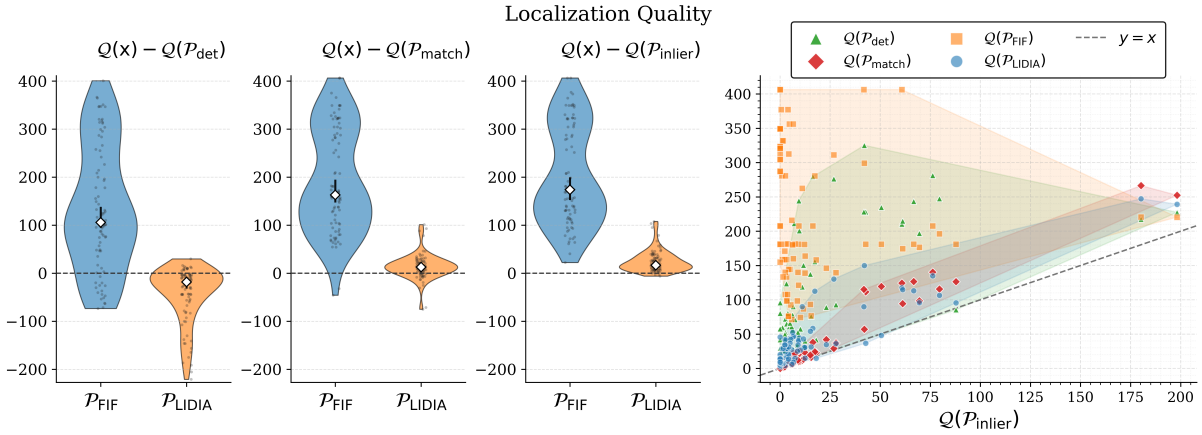

Fig. 6. **Closer range:** The left-most panels show the Information Gain (IG) difference between the predicted estimates and the detected, matched, and actual IG, respectively. The right-most panel shows IG as a function of the actual IG.

on the assumption that directly illuminated landmarks will yield more stable detections, and the results indicate that this is actually the case, since LIDIA remains a faithful representation for matched and actual IG. Finally, the right part of Table I shows the importance of knowing where to look when illumination conditions are not favorable. The difference between FIF and LIDIA in the localization accuracy is even greater in this scenario compared to the Database range experiment. We can see that LIDIA, when possible, chooses to look towards illuminated and informative regions of the observed spacecraft, resulting in significantly better localization accuracy.

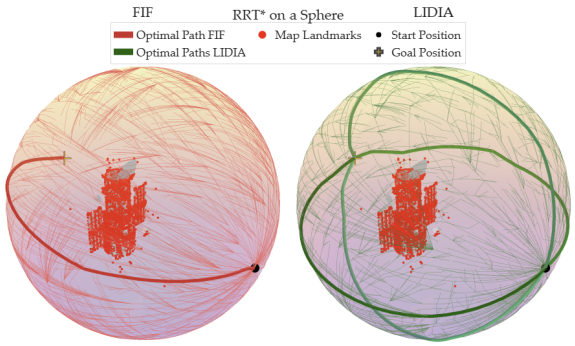

Fig. 7. The planned trajectories of RRT* under different lighting conditions for FIF (left) and LIDIA (right).

C. Qualitative Planning Result in Space Scene

We simulate a trajectory design experiment, where the camera can move on a sphere around the observed spacecraft. We assume a center-pointing trajectory that maintains the full spacecraft in view and seek to maximize the total localization quality along the trajectory, while minimizing the total trajectory arclength. The start and goal positions are initialized along the positive and negative directions of the world-frame e_{y} axis, respectively. Specifically, the initial state is placed along $+e_{\mathrm{y}}$, while the terminal state is symmetrically positioned along $-e_{\mathrm{y}}$. In contrast, the Sun illumination directions are aligned with the remaining principal axes of the world frame, namely $\pm e_{\mathrm{x}}$ and/or $\pm e_{\mathrm{z}}$, such that the lighting configuration is orthogonal to the start–goal displacement axis.

For a path of camera poses $\Lambda = \{T_{\mathrm{cw}}^{(0)}, \ldots, T_{\mathrm{cw}}^{(K)}\}$, the total cost of the path is defined:

$$\mathcal{J}(\Lambda) = \frac{\sum_{k=0}^{K-1} \|t_{\mathrm{wc}}^{\mathrm{w}(k)}\| \cos^{-1}(\hat{t}_{\mathrm{wc}}^{\mathrm{w}(k)} \hat{t}_{\mathrm{wc}}^{\mathrm{w}(k)})}{\epsilon + \sum_{k=0}^{K} \mathcal{Q}(\mathcal{P}(T_{\mathrm{cw}}^{(k)}))} \tag{31}$$

Before running the planning experiments, we intentionally moved the sphere along the $+e_{\mathrm{x}}$ axis creating an imbalance in the information gain calculations. Since we select the same random nodes across all trials, this displacement should bias FIF to always select the same path across all trials.

Fig. 7 shows the selected paths of RRT* for each of the different Sun directions. As we can see, LIDIA adapts based on the different Sun directions and selects plans that look towards illuminated regions of the spacecraft, even though other regions are closer and more informative. As a result, LIDIA can be used to avoid selecting paths that lead to completely dark images by taking into account the sun direction.

VII. Conclusions

This work addresses the challenge of predicting the localization quality of a specified pose for Perception-Aware Planning. We introduced LIDIA, a framework that integrates illumination awareness to estimate which landmarks will contribute to pose estimation. Our results demonstrated a reliable prediction of localization quality without requiring direct access to the captured image. In conclusion, our work highlighted the importance of taking into account the lighting conditions of the environment in perception-aware planning.

Incorporating illumination awareness enabled the development of a faithful representation of the perception model, which consistently improved the robustness of the planning module. The seamless integration of photometric information, without relying on costly rendering, opens new avenues

for perception-aware planning, particularly in environments where visual localization is prone to failure, ultimately advancing the frontier of robotic autonomy.

An important direction for future work is to investigate the algorithmic and architectural modifications required to achieve real-time performance, including computational simplifications, structured approximations, and implementations suitable for parallelization and onboard deployment. Moreover, the impact of more sophisticated illumination models that better capture real-world lighting phenomena, such as cast shadows, inter-reflections, and sensor-related artifacts, needs to be evaluated. Finally, real-world experiments can assess robustness, generalizability, and practical applicability under realistic sensing and environmental conditions.

REFERENCES

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5297–5307.

[2] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1808–1817.

[3] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in International Conference on Computer Vision (ICCV). IEEE, 2015, pp. 2938–2946.

[4] M. Ding, Z. Wang, J. Sun, J. Shi, and P. Luo, "Camnet: Coarse-to-fine retrieval for camera re-localization," in International Conference on Computer Vision (ICCV), 2019, pp. 2871–2880.

[5] S. Chen, X. Li, Z. Wang, and V. A. Prisacariu, "Dfnet: Enhance absolute pose regression with direct feature matching," in European conference on computer vision (ECCV). Springer, 2022, pp. 1–17.

[6] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," in Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3302–3312.

[7] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6684–6692.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." ACM Trans. Graph., vol. 42, no. 4, pp. 139–1, 2023.

[10] M. Pietrantoni, G. Csurka, M. Humenberger, and T. Sattler, "Self-supervised learning of neural implicit feature fields for camera pose refinement," in International Conference on 3D Vision (3DV). IEEE, 2024, pp. 484–494.

[11] P. Jiang, N. Bendapudi, S. Saripalli, and G. Pandey, "3dgs-loc: 3d gaussian splatting for map representation and visual localization," Journal of Autonomous Vehicles and Systems, vol. 5, no. 3, p. 031004, 2025.

[12] S. Se, D. Lowe, and J. Little, "Global localization using distinctive visual features," in International Conference on Intelligent Robots and Systems (IROS), vol. 1. IEEE, 2002, pp. 226–231.

[13] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2009, pp. 2599–2606.

[14] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7199–7209.

[15] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in Conference on Computer Vision and Pattern Recxognition (CVPR), 2019.

[16] A. R. Cassandra, L. P. Kaelbling, and J. A. Kurien, "Acting under uncertainty: Discrete bayesian models for mobile-robot navigation," in International Conference on Intelligent Robots and Systems (IROS), vol. 2. IEEE, 1996, pp. 963–972.

[17] G. L. Mariottini and S. I. Roumeliotis, "Active vision-based robot localization and navigation in a visual memory," in International Conference on Robotics and (ICRA). IEEE, 2011, pp. 6192–6198.

[18] Z. Zhang and D. Scaramuzza, "Perception-aware receding horizon navigation for mavs," in International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 2534–2541.

[19] L. Bartolomei, L. Teixeira, and M. Chli, "Semantic-aware active perception for uavs using deep reinforcement learning," in International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021, pp. 3101–3108.

[20] Z. Zhang and D. Scaramuzza, "Fisher information field: an efficient and differentiable map for perception-aware planning," arXiv preprint arXiv:2008.03324, 2020.

[21] L. Di Giammarino, B. Sun, G. Grisetti, M. Pollefeys, H. Blum, and D. Barath, "Learning where to look: Self-supervised viewpoint selection for active localization using geometrical information," in European conference on computer vision (ECCV). Springer, 2024, pp. 188–205.

[22] J. Chen, Y. Wang, P. Hou, X. Chen, and Y. Shao, "Dark-slam: A robust visual simultaneous localization and mapping pipeline for an unmanned driving vehicle in a dark night environment," Drones, vol. 8, no. 8, p. 390, 2024.

[23] J. Li, Z. Kuang, G. Lu, Y. Peng, W. Shang, J. Li, and W. Wei, "A lightweight stereo visual odometry system for navigation of autonomous vehicles in low-light conditions," Wireless Communications and Mobile Computing, vol. 2022, no. 1, p. 5249449, 2022.

[24] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," IEEE Transactions on Image Processing, vol. 30, pp. 2340–2349, 2021.

[25] J. Gao, C. Gu, Y. Lin, Z. Li, H. Zhu, X. Cao, L. Zhang, and Y. Yao, "Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing," in European conference on computer vision (ECCV). Springer, 2024, pp. 73–89.

[26] T. Zhang, K. Huang, W. Zhi, and M. Johnson-Roberson, "Darkgs: Learning neural illumination and 3d gaussians relighting for robotic exploration in the dark," in International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 12 864–12 871.

[27] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge university press, 2003.

[28] Y. Wu and Z. Hu, "Pnp problem revisited," Journal of Mathematical Imaging and Vision, vol. 24, pp. 131–141, 2006.

[29] P. Sturm, "Pinhole camera model," in Computer Vision: A Reference Guide. Springer, 2021, pp. 983–986.

[30] R. Hartley, Multiple view geometry in computer vision. Cambridge university press, 2003, vol. 665.

[31] K. M. Frey, T. J. Steiner, and J. P. How, "Towards online observability-aware trajectory optimization for landmark-based estimators," arXiv preprint arXiv:1908.03790, 2019.

[32] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[33] B. O. Community, Blender - a 3D modelling and rendering package, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[34] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Conference on Computer Vision and Pattern Recognition (CVPR) workshops, 2018, pp. 224–236.

[35] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4938–4947.

[36] National Aeronautics and Space Administration, "Hubble space telescope 3d model," https://science.nasa.gov/resource/hubble-space-telescope-3d-model/, 2023, accessed: 2025-09-06.

[37] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic et al., "Benchmarking 6dof outdoor visual localization in changing conditions," in Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8601–8610.